WHITE PAPER

**Contents**

# PLANNING CONSIDERATIONS FOR GEOGRAPHICALLY DISPERSED CLUSTERS USING IBM HACMP/XD: HAGEO TECHNOLOGY

## ABSTRACT

Developing a highly available environment requires skillful planning and preparation. Prior to executing any systems integration strategy, every aspect of the project, from definition and design to implementation and testing, must meet your business objectives. The only way to achieve a successful implementation is to plan for it. The audience for this white paper is Information Technology professionals who have a general understanding of IBM HACMP™ for AIX 5L™ and geographically dispersed clusters. This paper will walk you through the factors you should consider when planning for your implementation of IBM HACMP/XD: HAGEO Technology.

# Introduction

The success of your business depends on the availability of your applications, networks and data. Information must flow freely to wherever it is needed in the company, whenever it is needed. As an IT professional, your challenge is to keep your business operational by keeping information systems available 24/7. Downtime in any form – planned or unplanned – could cost your company lost revenue, lost opportunity and even customer loyalty.

Any number of failures could cause an unplanned outage. Something as dramatic as an earthquake or hurricane could devastate your data center, or something as mundane as a power outage or network adapter failure could take down your servers. The fact remains that every minute of downtime leaves your employees idle and your customers worried. What you need is a well-planned solution that ensures that critical data and applications remain available, no matter what.

## Protecting Your Applications and Data from Unplanned Downtime

IBM addresses your need for around the clock availability by offering world-class solutions for achieving long distance mirroring, fallover and resynchronization of data during site recovery. IBM continues to set the standard for availability with its High Availability Cluster Multiprocessing (HACMP) XD (*eXtended Distance*) set of solutions. The HACMP/XD feature provides two distinct software solutions for disaster recovery. Added to the base HACMP for AIX 5L software, they each enable a cluster to operate over extended distances at two sites.

**HACMP/XD: HAGEO Technology** provides unlimited distance data mirroring. It is based on the IBM High Availability Geographic Cluster for AIX 5L (HAGEO) v 2.4 product. HAGEO Technology extends an HACMP for AIX 5L cluster to encompass two physically separate data centers. Data entered at one site is sent across a point-to-point TCP/IP network and mirrored at a second, geographically distant location.

**HACMP/XD: Remote Copy** increases data availability for IBM TotalStorage® Enterprise Storage Server® (ESS) volumes that use Peer-to-Peer Remote Copy (PPRC) to copy data to a remote site for disaster recovery purposes. HACMP Remote Copy takes advantage of the PPRC fallover/fallback functions and HACMP for AIX 5L cluster management to reduce downtime and recovery time during disaster recovery. This provides a shorter distance hardware-based solution.[1]

[1] Refer to the IBM/Availant White paper titled "Automated Recovery Management with HACMP/XD and PPRC" for more information on this solution at http://www.ibm.com/servers/eserver/pseries/software/whitepapers/hacmp_pprc.pdf

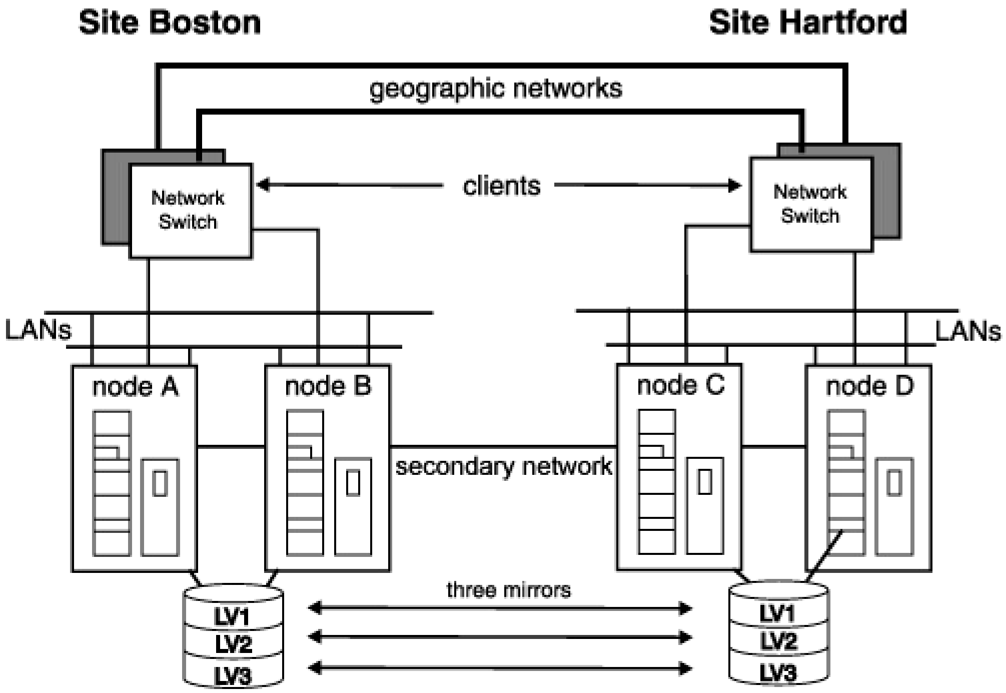
Besides the solutions offered by HACMP/XD Remote Copy and HAGEO Technology, storage devices in a Storage Area Network (SAN) can also provide disaster recovery. A Storage Area Network is a high-speed network that allows connections between storage devices within the distance supported by Fibre Channel. Geographically separated storage devices and servers can utilize AIX 5L Logical Volume Manager (LVM) mirroring to replicate data at geographically separated sites.

This white paper will help you examine your business requirements so that you can decide if an availability plan that includes geographically dispersed clusters is right for you. We will walk you through the considerations you should make when implementing HACMP/XD: HAGEO Technology. In the process, we sometimes suggest that given certain factors, you might be better served with the HACMP/XD: Remote Copy solution.

## What is HACMP/XD: HAGEO Technology?

The HACMP/XD: HAGEO Technology environment is a distributed cluster that spans two sites separated by enough distance to prevent the same disruption from disabling both sites. Redundant geographic point-to-point TCP/IP networks connect the sites, providing recovery over any distance.

The following is an example diagram of an HACMP/XD: HAGEO Technology cluster:



Each location maintains a real-time mirror of your business-critical applications and data so that when one site encounters a disruption, HACMP/XD: HAGEO Technology responds by providing automatic fallover and resynchronization. This automation feature of HACMP/XD: HAGEO Technology ensures that your system is up and running in a significantly shorter time frame than if you needed to rely on loading back up tapes at a second site. Once HACMP/XD: HAGEO Technology is configured for your environment, hours of travel and heads-down system administration turn into just minutes of monitoring HACMP/XD: HAGEO Technology while it brings resources back online. And because the transition is automated, you have the opportunity to capture best practices, reducing the risk of user error. Best of all, your users encounter very little down time before they are able to access up-to-date data and resume normal business operations.

Solid planning is a key determinant to deliver the benefits of a geographically dispersed cluster. Many aspects must be considered, such as business requirements, sites, servers, disk subsystems, and networks. The goal to provide continuous availability and optimal performance must be balanced with cost, performance, and availability.

## Is HACMP/XD: HAGEO Technology Right for You?

HACMP/XD: HAGEO Technology is a powerful solution that can provide significant cost savings through reduced downtime in the event of a system or site failure. Used properly, it can show a considerable return on investment by preventing data loss and minimizing application outages. However, it is not the appropriate solution for all environments. The first step is determining if HACMP/XD: HAGEO Technology is right for your business.

## Categorize Your Applications

Deciding on a geographically dispersed cluster solution starts with identifying the top applications at risk in your environment and defining their availability requirements. Divide your applications into these two categories:

- **Business or revenue-generating applications**: Applications that need to be available to generate revenue and increase market share, such as e-commerce storefronts, banking applications, or quoting and ordering systems.

- **Operational applications**: Applications that you need to maintain your business on a day-to-day basis, such as CRM or accounting systems.

Business or revenue-generating applications present the biggest risk to your company if they were unavailable for any length of time. Therefore, you need to protect these applications by ensuring that a disaster in your system environment won't prevent your users from accessing up-to-date data they need to keep generating business. Although the operational applications are needed to maintain the business, by themselves they typically do not warrant the cost associated with implementing a geographically dispersed cluster solution.

## Determining Your Business Requirements

To determine which level of availability meets your needs, you should consider the following questions:

### What is the financial impact of downtime to your organization?

Consider what happens when your users and customers cannot access your system. How many online transactions are your customers unable to complete? What is the overall impact of the customer service your company is unable to provide?

### How long can you operate without your business applications?

If your business can't afford to be down for the amount of time it takes for your team to travel to your backup site and manually bring your resources back online, you would benefit from the automated fallover and recovery features of HACMP/XD: HAGEO Technology. Additionally, automated features reduce user errors and provide the capability to transition workloads or resources for site or node maintenance.

### How much data can you afford to lose?

If you need to guarantee access to real-time data before and after a failure, you need a solution like HACMP/XD: HAGEO Technology that provides automated synchronous data replication and resynchronization capabilities. If you can tolerate a certain amount of lost data, you can take advantage of HACMP/XD: HAGEO Technology's asynchronous capability, which provides improved performance over its synchronous mirroring mode. If you require high performance with no loss of data, the HACMP/XD: Remote Copy with PPRC and ESS hardware-based replication may be a better solution for you.

### How far apart are your two sites?

Some geographically dispersed clusters require a fiber optic connection between the primary and backup sites, which can impose a constraint on the distance permitted between sites. HACMP/XD: HAGEO Technology does not require a fiber connection, and has no distance limitations.

### How key is I/O performance to your environment?

Implementing higher levels of availability typically has implications for overall system performance. If I/O performance is critical, then it is important to adequately prepare for the additional performance load of HACMP/XD: HAGEO Technology. If you require very high performance, and the sites meet the distance requirements, the HACMP/XD: Remote Copy with PPRC and ESS hardware-based replication may be a better solution for you.

### What are the read and write characteristics of the application(s) that will be using this system?

HACMP/XD: HAGEO Technology provides consistent performance for workloads with a balanced level of read/write operations and for read-intensive workloads. All write operations are transferred across the network, so an application that is extremely write-intensive may not perform well. Read-intensive environments that have infrequent large write requests may improve consistent performance by using asynchronous writes, because the pending read requests are not delayed. Applications that flood data to disk sporadically will require larger bandwidth in the networks that connect the sites.

## Business Impact Analysis

Before you determine the details of your system, you should have a clear understanding of why you need HACMP/XD: HAGEO Technology. You need to define your company vulnerabilities, and plan how a geographically dispersed cluster can address these vulnerabilities to protect your business.

The first step is to identify the business goals you need to meet. This will help you decide whether the solution you design will have real business value. Start by answering the following question:

Why are you implementing this system?
*Examples of how you may answer this are:*

- To maintain up-time of nearly 100% for business critical applications

- To ensure that our applications are accessible in a matter of minutes instead of days after a disaster.

- To protect critical business applications and data in the event of a site disaster.

## Business Vulnerability

It is important to understand the threats your system faces and the impact that any of these events would have on your business. Determine the likely consequences of an unplanned and unwanted event so that you can calculate the business impact of a possible outage and the importance of avoiding it. Business impacts typically fall into one of three categories:

- **Direct impacts:** The actual value of the capital assets that are lost

- **Indirect impacts:** The damages caused by the business' inability to operate after losing assets

- **Consequential impacts:** Losses that result from not being able to do business while recovering from a disaster

Performing a business impact analysis will position you to understand the threats to your business and how HACMP/XD: HAGEO Technology can help protect your business. HACMP/XD: HAGEO Technology addresses the indirect and consequential impacts, which are often the most damaging. Lost physical equipment may be expensive to replace but lost business information can be impossible to replace.

# Gathering Information for Your Design

In this section we will walk you through the areas for which you need to gather information. In the section that follows, we will then take into account possible answers to the questions raised in this section.

## Single Points of Failure Analysis

The major goal throughout the geographic cluster planning process is to eliminate Single Points of Failure (SPOF). A single point of failure exists when a critical cluster function is provided by a single component. Through good planning, you can eliminate the single points of failure and provide the best possible performance.

As you design your geographically dispersed cluster, identify and address all potential single points of failure. Cluster components that are potential single points of failure include:

- Power sources
- Network switching equipment
- Sites
- Nodes (Servers)
- Geographic networks and adapters
- Local area networks and adapters
- Storage and storage adapters
- Applications

How will application services be affected if one of these components were to become unavailable?

## Geographical Topology

A second goal of the planning process is to design the geographic topology and fallover strategy that meet your business requirements. The following components need to be planned and defined. We provide questions for each to help you gather information relevant to your geographically dispersed cluster.

### Sites

Begin planning the cluster by thinking about the cluster as a whole and how each site relates to the other. Consider the following:

- Do you already have a second site in mind or are you planning to build one?
- Which applications are running at each site?
- What applications will you include in the cluster?
- Where are the users located?

- How much processing power do you need at each site?
- What type of backup communications systems do you plan to use for heartbeat traffic? IP, serial link, or modem?

## Nodes

After determining your site characteristics, you need to consider the node characteristics:

- How many servers (nodes) exist?
- How many nodes will exist at each site?
- What are your local and remote failure requirements?
- What applications are running on each node?
- What applications will fallover to the remote site?
- What networks are configured on each node?
- What disk/logical volumes/filesystems are defined on each node?
- What will each node do when another node fails?

## Storage

Storage technology and how it is accessed by both the system and applications need to be considered.

- What type of storage subsystems do you have?
- Does your storage subsystem provide caching ability?
- What type of storage are you planning for the backup site?
- What is your data made up of? Raw logical volumes, Journal File Systems (JFS) or both?
- Is all your data located on external disks?
- How do you plan to perform the initial synchronization of data?

### Applications

Understanding your application is critical to normal operations and recovery.

- What are your applications?
- What functions do they perform?
- What are the read/write characteristics of your applications?
- Where is the application software (binaries) installed, internal or external disks?
- Can your applications be recovered without manual intervention?
- Can your applications survive a system crash?
- What types of users (customers, business partners, and staff) access the applications?
- How will users re-connect after a site fallover?
- Do your applications have a hostname dependency?
- Do your applications have an IP-label or MAC address dependency?
- Is your software licensing tied to a specific machine or CPU?

- What are the procedures for starting, stopping and recovering your applications?
- Are there any requirements for application development in the production environment?

## Inter-site Network Bandwidth Requirements

The geographic *primary* networks interconnect the primary and backup sites, and are dedicated networks, used only for mirroring data. Performance of a geographic mirror is based on the bandwidth and latency of these communication links.

To measure bandwidth requirements, start by identifying the logical volumes that contain the critical data to be mirrored. Measure the write I/O throughput and performance for these logical volumes. Use the data captured to help size and cost the geographic networks. The gmdsizing tool or the AIX 5L filemon command can be used to gather the necessary disk I/O information.

The **gmdsizing** tool monitors disk utilization over a given period of time and prints a report. This report can then be used to help determine bandwidth needs.

The **gmdsizing** tool is available from the following sources:

- Download from the web, http://www-1.ibm.com/servers/eserver/pseries/solutions/ha/apps_license.html
- The HACMP for AIX 5L installation media
- The HACMP/XD: HAGEO Technology installation media

All three sources contain a README file with detailed instructions for using the tool and interpreting the output.

Example:

```
gmdsizing -i10 -t60 -v vg3 -V -f
/tmp/lv_gmdsizing_$(date +%Y%m%d%T)
```

The **filemon** command uses the AIX 5L trace facility to collect I/O information. A report is generated providing performance statistics for files, logical volumes, and physical volumes. The **filemon** command is part of the bos.perf.tools fileset, which is installable from the AIX 5L base installation media. Reference the *AIX 5L Performance Tools Handbook* for detailed information on using **filemon**.

Example:

```
filemon -o /tmp/lv_filemon_$(date +%Y%m%d%T)
-O lv,pv;sleep 600;trcstop
```

Disk utilization should be collected for a time interval when peak processing occurs. If the system is running a nine to five operation, measuring disk utilization in the middle of the night will not yield meaningful information. Similarly, measuring over a very short period of time is not likely to yield representative data. You need to understand how a workload varies over time. This typically includes:

- Busy periods during a day / week or month
- Year end / end of quarter processing
- Overnight batch processing

It is better to run the tools over a longer period of time than a shorter one, so as to capture peaks and troughs. When specifying the observation interval however, it is better to keep this larger rather than smaller. One line of data will be written per disk per interval so a very large amount of data will be collected if you have a small interval and/or a large number of disks. Remember that the more data you collect, the more data you will have to process.

If the system you are measuring has local LVM mirroring enabled be sure to take this into account when selecting what to measure. For example, if you have two-copy mirroring enabled for the logical volumes, one logical write from the application generates two physical writes to the disk devices. Rather than selecting an entire volume group to be monitored, select just those disks that contain one copy of the mirrors. If your volume group is laid out such that you cannot easily do this, then remember that you are potentially recording twice as much write activity as the application is generating, which should be factored out of your data analysis.

# Analyzing Gathered Data to Drive Your Design

Once you have gathered data describing the characteristics of your application environment and your availability goals, it is time to use the information to drive your design. This section provides possible answers to the questions raised in the previous section. Note that it would be impossible to address all possible answers to the questions raised in this paper, but the following sub-sections will give you the basis for planning what is required to successfully implement HACMP/XD. Also realize that planning of a cluster is an iterative process. Continued refinement of a design is expected as you clarify collected data.

Performance is a critical issue in the overall success of the project. Use this phase as an opportunity to identify potential tuning opportunities and initiate action to resolve issues now. Any performance issue identified as a result of gathering and analyzing data should be addressed prior to deploying an HACMP/XD HAGEO Technology solution. For example, systems that are performing at a high utilization and have limited resources available are not good candidates for a HACMP/XD HAGEO Technology implementation. Addressing limited resource availability (network bandwidth, memory, storage space, for example) would be required before implementing HACMP/XD HAGEO Technology.

## Sites

Selection of the sites is a critical factor. The physical location of the primary site should be fundamentally stable in order to minimize the likelihood of a major service interruption. When considering a secondary site, it must be capable of supporting the same critical business functions as the primary site.

The selection of a site also depends on the location of the clients and the accessibility of the site. If most of the clients are located at the primary site and the physical access to the secondary site is restricted, plan for one of the sites to be the active (or primary) site and the other to be the hot standby (or backup) site. If clients are located at both sites, consider a mutual takeover configuration where the first site runs one group of applications and the second site runs another group of applications, with each site backing up the other. For this mutual takeover design, be sure to consider the performance issues of the client network and the server capacity when all applications reside at the same site, in the event of a disaster.

When determining the distance between sites, keep in mind that increasing the distance between sites decreases the chances that both sites will experience a disaster at the same time. However, the farther apart the sites are, the more complicated and expensive it is to have network connections between them to support sufficient bandwidth for your applications. Increasing the distance also increases network latency, which in turn impacts I/O performance.

## Nodes

It is important to understand the hardware configuration in order to verify that it is supported and capable of meeting required expectations. The server configuration must be capable of sustaining production in the event of either a local server failure or of an entire site failure. The software levels for the operating system and application are required to be at the same version and maintenance levels on all nodes.

For discussion purposes, consider three popular HACMP/XD: HAGEO Technology designs:

- Two nodes at each site
- Two nodes at the primary site and one node at the backup site
- One node at each site

### Two nodes at each site

The primary and the backup sites have full local takeover capability. All component failures are handled locally at each site. Only site failures or planned site maintenance result in fallover across sites. This configuration is preferred if you will have applications active at both sites.

Having multiple nodes at the backup site allows for maintenance to be performed to either server at that site, while still providing the backup capabilities for the primary site. This is possible because the workload can be moved to the backup node at the backup site, allowing it to continue to receive the updated data from the primary site and node.

### Two nodes at the primary site and one node at the backup site

All component failures are handled locally only at the primary site. The secondary site elevates any node failure to a site failure. Even though the backup site can be configured with its own application that would fallover to the primary, this design is more satisfactory if there are no applications running at the backup site. If applications were running on the backup, then maintenance on the backup site would necessitate inter-site resource takeover.

This configuration is supported, but you should remember that if maintenance is being performed on the backup site, the node at the backup site will be down and the data updated at the primary site cannot be synchronized to the backup site. When the node at the backup site comes up, the primary site has to synchronize all changes that occurred while the backup site was down.

### One node at each site

This design is more restrictive. It may handle disk or adapter failures locally, but node failures are propagated to site failures, because there are no local peer nodes available in this model.

## Networks- primary

Planning the network component is among the most important tasks for the HACMP/XD: HAGEO Technology solution. Providing the appropriate bandwidth will minimize performance impact to your applications. HACMP/XD: HAGEO Technology supports multiple networks for the purpose of increasing availability and throughput.

*Geographic primary networks* carry the mirrored data and HACMP heartbeat between sites. Plan for one or more independent point-to-point networks running TCP/IP for regular service between sites. To decide how many networks are needed, take into account the projected workload of your sites. HACMP for AIX 5L clients should not access the primary networks directly.

It is highly recommended that the HACMP/XD: HAGEO Technology solution be configured with dedicated IP networks for the geographic primary networks.

In planning for HACMP/XD: HAGEO Technology capacity you need to determine the network bandwidth, network latency and disk latency. This is an iterative process because bandwidth required for the data may have to be increased to compensate for network latency.

### Hints on determining Network Bandwidth Requirements

The following example demonstrates how you would calculate the I/O load of your application. This example is included for discussion purposes only. Remember it is important to capture data for the peak times the system is utilized in order to plan for proper capacity.

To determine the minimum required network bandwidth in this example the following **filemon** command was used. For this example, a collection period of only 60 seconds was used, but for capturing data on your system, you should use a longer interval around peak periods to obtain more accurate information.

```
filemon -o /tmp/lv_filemon_$(date +%Y%m%d%T)
-O lv;sleep 60;trcstop
```

The output from the **filemon** command shows the following logical volumes in the most active list. Logical volume lvr3s2 will be used in this example as one that will be mirrored by HACMP/XD: HAGEO Technology.

```
Most Active Logical Volumes
---------------------------------------------
util  #rblk  #wblk   KB/s  volume description
---------------------------------------------

 0.04  13680   5856  160.7  /dev/lvr3s2  raw

 0.04  13632   5856  160.3  /dev/lvr3m2  raw

 0.04  13664   5856  160.5  /dev/lvr3m1  raw

 0.04  13648   5856  160.4  /dev/lvr3s1   raw
```

To calculate bandwidth for logical volume lvr3s2 writes:

#wblk * 512 / 1024 = KB (Kilobytes)/interval

5856 * 512 / 1024 = 2928 KB/interval

2928 * 8 = 23424 Kb (Kilobits)/interval

23424/60 = 390 Kb/sec

It is recommended that the HACMP/XD: HAGEO Technology and network overhead do not exceed 75% utilization.

bandwidth = Kb/sec / .75

390 / .75 = 520 Kb/sec or .5 Mb/sec

In this case a minimum network bandwidth of .5 Mb/sec would be required.

The **filemon** output also supplies more detailed information on write sizes for each logical volume:

```
-----------------------------------------------------
Detailed Logical Volume Stats   (512 byte blocks)
-----------------------------------------------------
```

```
VOLUME: /dev/lvr3s2  description: raw

reads:                  1710    (0 errs)

read sizes (blks): avg  8.0 min      8 max      8 sdev  0.0

read times (msec): avg 0.766 min 0.362 max 26.199 sdev 1.853

read sequences:       1710

read seq. lengths: avg   8.0 min       8 max   8  sdev  0.0


writes:                 732     (0 errs)

write sizes (blks): avg   8.0 min    8 max       8 sdev  0.0

write times(msec): avg  1.411 min 0.972 max  27.106 sdev 1.556

write sequences:      732

write seq. lengths:  avg   8.0 min    8 max    8 sdev    0.0
```

This shows 732 writes per interval with the average of 8 blocks per write when writing locally.

In order to determine the average write time in a HACMP/XD: HAGEO Technology environment, we need to know the time to transfer the data over the network, the network latency and disk latency at the local and remote sites.

The data transfer time can be calculated as follows from the **filemon** report. The report shows the average number of blocks written. HACMP/XD: HAGEO Technology requires an additional block for the acknowledgement of the written data.

data transfer time = average blocks +1 * 512 *8 / network bandwidth

(8 + 1) * 512 * 8 = 36864 bits / .5 Mb/sec = **74 ms**

Network latency can be measured if the network is available, a **ping** test will provide some idea of the round trip delay. If the network is not available, network providers should be able to supply information on network latency. For this example a network latency of 17 ms was used. The network latency will never be less than 1 ms per100 miles, the latency added by network hardware and overhead adds to the total network latency.

latency = **17 ms** for this example

The average disk latency can be determined from the **filemon** report by using the average write time. To account for the write time at the local and remote sites this value is multiplied by 2.

disk latency = average write time * 2

1.4 ms * 2 = **2.8 ms**

The total write time in this example would be the following:

write time = data transfer time + network latency + disk latency

74 ms + 17 ms + 3 ms = 94 ms

For a synchronous geographic mirror this would limit the writes to the following:

Writes per interval = interval / write time

60 / .094 = 638

This example shows the write rate of 638 per interval is less than the 732 per interval reported by **filemon**. This decreased rate could impact application response time. Additional network bandwidth would be required to compensate for the addition of network latency. In this example, if the network bandwidth were increased to 1 Mb/sec then a theoretical write rate of 968 could be maintained.

The following table shows expected throughput for various networks:

| **Name** | **Bandwidth** |
|---|---|
| Ethernet | 10/100/1000 Mb/sec |
| T1 | 1.544 Mb/sec |
| E1 | |

| | |
|---|---|
| | 2 Mb/sec |
| E3 | 34 Mb/sec |
| T3 | 45 Mb/sec |
| OC3 (STM1) | 150 Mb/sec |
| ATM | 155 Mb/sec |
| Fibre Channel | 1 or 2 Gb/sec |

## Networks – secondary

G*eographic secondary networks* carry the HACMP heartbeats (cluster health messages) between sites, just as the serial network does for HACMP for AIX 5L. To ensure that HACMP for AIX 5L can detect a global network failure and handle this event gracefully and quickly you must include a secondary communications path in your HACMP for AIX 5L configuration. This secondary network must be defined to HACMP as a network for heartbeat traffic only. You can use an IP network or a dedicated phone line. If you set up an IP network for heartbeat traffic, you may need to tune the heartbeat values to allow for the longer distance.

If you use the fault-tolerant public switched telephone network to communicate point-to-point between the sites, you must enable the Dial Back Fail Safe (DBFS) option when you configure the secondary HAGEO networks. When communication is not possible over the regular networks (site isolation), the non-dominant site nodes use DBFS to dial the dominant site nodes. If any dominant site nodes are available, all the nodes at the non-dominant site are shut down. Only nodes at the dominant site remain up in the event of site isolation to prevent data divergence. If the dominant site does not answer, the calling site assumes that a disaster has occurred and initiates the takeover process.

## Geographic Mirroring

### Asynchronous versus Synchronous

When determining the right mirroring technology for your business, first define your recovery objectives.

*Recovery Time Objective (RTO)* defines the requirement for how quickly a business and applications must be restored following an operational interruption. How long can you afford to be without your systems? *Recovery Point Objective (RPO)* defines the required point in time to which information must be restored. How much data can you afford to recreate?

In the HACMP/XD: HAGEO Technology implementation, data replication can be performed in one of three modes: asynchronous, synchronous or synchronous with mirror write consistency (MWC). In synchronous mode data is written to the secondary site first and then to the primary site. The write does not return to the application until it completes at both sites. When the write returns, data is known to be secure on both the primary and secondary sites. In MWC extra logging to the state maps is added to synchronous mode, and MWC mode allows the local and remote writes to overlap. In asynchronous mode the write completes immediately on the primary site and the secondary site is updated when time allows.

All modes have their advantages. Synchronous mode has an advantage in disaster recovery where data loss cannot be tolerated. Asynchronous mode has an advantage in write-intensive environments where performance is a higher priority than prevention of data loss. MWC is the best choice when you require reasonable performance and guaranteed no data loss in a disaster. Defining your RTO and RPO will help you determine which option is the better option for you.

Utilizing disk cache reduces disk I/O time and improves geographic mirroring performance considerably. For synchronous mirrors, priority should be given to caching local and remote data logical volumes. For asynchronous and MWC environments, priority should be given to caching local state maps and remote data volumes.

## Storage

Storage capacity must be sufficient to handle at least one mirror of all data to be protected from the other site. Multiple mirrors of the same data are recommended to eliminate disks as a single point of failure. This can be done by choosing a storage technology that supplies protection (such as RAID) or by utilizing the LVM feature of AIX 5L.

### Hints on Implementing Storage

An average disk has a seek time of 8-10 msec. When HACMP/XD: HAGEO Technology is involved there will not be one seek, but many. If you are using asynchronous mirroring or synchronous with mirror write consistency, then you will have seek activity for the data as well as for the state maps, which contain information used to track mirror status across sites. The associated seeks with writing the data will also have to occur at both sites. Seeks for state maps occur only on the 'writing' side. The key mechanism for improving disk response and reducing the latency is to make use of disk caches if at all possible. If the volumes in your disk subsystem are individual physical disks (JBOD configuration) it is preferable to put the state maps on separate disks than the data.

Writing large blocks of data is the most efficient use of HACMP/XD: HAGEO Technology. If you have a choice, use raw devices instead of JFS filesytems for the application. When writing to raw disk you are not restricted to the 4KB page limits of JFS.

## Applications

Once the business critical applications are identified, it is important to ensure that the application itself is configured for optimal availability and recovery. All application data must be stored on shared disks, so that they are available to the node taking over in the event of a failure.

Most applications require the use of a license key; HACMP for AIX 5L will have to be configured to handle the license key during startup, fallover and reintegration activities.

## Client Access

Another critical component that often gets overlooked when planning a geographically dispersed cluster is client access to the critical applications in the event of a site fallover. You must consider how to ensure that user access to the applications at the remote location is handled automatically.

Client access is typically done over an IP network through the use of an IP address or a resolvable IP network interface name. The client makes a request to the application and is automatically directed to the location and server without the client having to know (or care) about the physical address or location of the application.

Steps must be taken to address access to the remote location from both the physical and logical network configuration. Clients must have access to the physical network segment of the remote location and have access rights through firewalls and inclusion into any Virtual Private Networks (VPN) that make up the Wide Area Network (WAN). Because of the complexities that comprise most corporate networks, considerations for client access when planning geographically dispersed cluster environments is very important.

## Single Points of Failure

This section identifies potential single points of failure associated with an HACMP/XD: HAGEO Technology solution and describes possible solutions to eliminate them. In general, eliminating single points of failure locally is simpler, more efficient and less expensive.

### Power Sources

All nodes and subsystems at each site should be powered with redundant power supplies, powered by separate circuits.

### Network Switches

Using more than one switch avoids having a network switch be a single point of failure.

### Nodes

Configuring a cluster and defining fallover and reintegration responsibility between nodes and sites eliminates a node as a single point of failure.

### Geographic network and adapters

Each site should be connected to at least two geographic primary networks to avoid a SPOF. These networks should be used only for data replication traffic. No clients should have access to these networks. A geographic secondary network or a dedicated phone line is also necessary for heartbeat traffic, to detect site isolation.

### Local area networks (LANs) and adapters

Multiple LANs reduce the risk of an outage due to a network going down, while HACMP for AIX 5L handles the availability of the network interfaces as long as multiple network interface cards are installed in each node.

### Disks and adapters

To eliminate disks as a SPOF, all data should be protected either by LVM mirroring or by vendor supplied technology such as RAID. This includes protecting the root volume group on all nodes. Multiple adapters or paths to the disk should also be configured for each node, with appropriate solutions necessary to recover from a failure of a data path.

### Applications

If the application fails on one node, HACMP for AIX 5L can be configured to restart the application on the same node, or to move the application and its associated resources to a different node through resource monitoring.

## Summary

Integrating a high availability strategy into a production environment can be a complicated process, but skillful planning and preparation will ensure that your implementation will provide the level of availability that you need to keep your business running smoothly in the event of unplanned downtime. Carefully considering the issues outlined in this white paper will help you get the most out of your geographically dispersed cluster.

Once you have configured HACMP/XD: HAGEO Technology in your environment, your users and customers are ensured access to their applications and data if your primary site experiences a catastrophe. When one site encounters a disruption, HACMP/XD: HAGEO Technology responds by providing automatic fallover and resynchronization at your back up site, providing applications in a significantly shorter time frame than if you needed to travel to a standby site and reload backup tapes manually. And because this transition is automated, you have the opportunity to capture best practices, reducing the risk of user error.

# Customer Reference

***The Principal Financial Group (PFG)***

**Business Need:**
PFG was using tape backups for its disaster recovery solution and wanted to change to an automated system that would help reduce administrative costs. Its goal was to be self-recoverable and to offer high availability to key AIX 5L-based applications.

**Solution Implemented:**
In place of the previous tape storage solution, PFG now has two data centers that are 28 km apart and are connected by dark fiber and Dense Wave Division Multiplexing (DWDM) technology. This extends the production network that supports the customer's key AIX 5L-based applications. PFG splits its critical servers between the two locations and uses HACMP/XD: HAGEO Technology to provide high availability and disaster recovery. The server solution consists of seven clusters spread across 14 servers running Tivoli® Framework, Netview®, DB2® and Peregrine.

**Benefits of the solutions:**
HACMP/XD: HAGEO Technology allowed PFG to create an automated highly available platform for its open systems environment, while reducing administrative costs, recovery time, and overall downtime.

# References

## Publications:

*IBM's HACMP/XD for AIX 5L for HAGEO Technology Concepts and Facilities*, September 2003, SA22-7955-00

*IBM's HACMP/XD for AIX 5L for HAGEO Technology Planning & Administration Guide,* September 2003, SC23-4862-00

*IBM's Disaster Recovery using HAGEO and GeoRM, May 2000, SG24-2018-03*

*IBM's Configuring Highly Available Clusters*, October 2002, SG24-6845-01

*IBM's AIX 5L Performance Tools Handbook,* August 2003, SG24-6039-01

*Availant's Geographic High Availability White Paper*, August 1995

# About the Authors

## Chris Fallon

Chris Fallon is a Principal Consulting Engineer at Availant. He has served as an information technology consultant and instructor for over fifteen years. He specializes in open systems availability, database and storage solutions, and has implemented a wide range of cluster, enterprise SAN, data replication and backup products for Fortune 500 companies.

## Jim Dieffenbach

Jim Dieffenbach is a Principal Engineer in Technical Support at Availant. His HAGEO experience includes system integration, technical support and quality assurance.

## Jill Broyles

Jill Broyles is the former Founder, President and CEO of Prevail Technology. She is an Information Technology professional with over twenty years experience in high tech, including business strategy, systems and availability consulting. Jill has served as a board member on several IBM Business Partner Councils, and as an Advisory Board member for various IBM Business Partners.

# About Availant

Availant is a provider of professional and engineering services. Founded in 1989, the company's customers include more than half of Fortune 500 companies in the finance, telecommunications, retail, health care and transportation industries, where high availability and continuous access to enterprise information is essential to business operations. Offerings include SAN and storage assessment, planning, design and implementation; data migration; availability assessment; HACMP for AIX 5L cluster planning, design and implementation; validation testing; course development and delivery in AIX 5L, clustering and storage; and development services across UNIX® and Windows®. Availant is headquartered in Waltham, Massachusetts. For more information, call Availant at 888.94.AVAIL or visit the company's Web site at http://www.availant.com/.

IBM Corporation
Marketing Communications
Systems Group
Route 100
Somers, New York 10589

Produced in the United States of America
June 2004
All Rights Reserved

This document was developed for products and/or services offered in the United States. IBM may not offer the products, features, or services discussed in this document in other countries. The information may be subject to change without notice. Consult your local IBM business contact for information on the products, features and services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information concerning non-IBM products was obtained from the suppliers of these products.
Questions on the capabilities of the non-IBM products should be addressed with the suppliers.

All performance information was determined in a controlled environment. Actual results may vary. Performance information is provided "AS IS" and no warranties or guarantees are expressed or implied by IBM.

The IBM home page on the Internet can be found at http://www.ibm.com.

The IBM high availability home page on the Internet can be found at www.ibm.com/servers/eserver/pseries/solutions/ha .